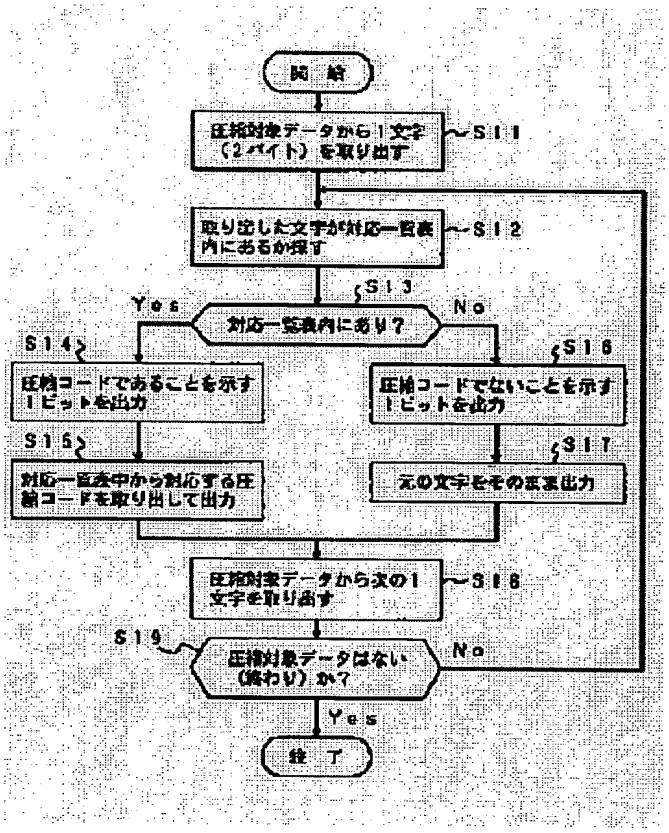
METHOD AND DEVICE FOR DATA COMPRESSION

Patent number: JP9069785
Publication date: 1997-03-11
Inventor: TODA ATSUKO
Applicant: TOSHIBA CORP
Classification:
- international: H03M7/42
- european:
Application number: JP19950222155 19950830
Priority number(s):

Abstract of **JP9069785**

PROBLEM TO BE SOLVED: To realize effective data compression in the case of taking Japanese- language text data as the compression object.
SOLUTION: A correspondence list where code data of specific characters and their compressed code data correspond to each other is prepared, and it is retrieved whether character code data of one character taken out from compression object data exists in the correspondence list or not (S11 and S12). If it exists in the list as the result (S13), its compressed code data is read from the correspondence list and is outputted (S14 and S15); but if it doesn't exist there (S13), data is outputted as it is (S16 and S17).



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－69785

(43)公開日　平成９年(1997)３月11日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｈ０３Ｍ　7/42		9382－5K	Ｈ０３Ｍ　7/42	

審査請求　未請求　請求項の数４　ＯＬ　（全 5 頁）

(21)出願番号　特願平7－222155

(22)出願日　平成７年(1995)８月30日

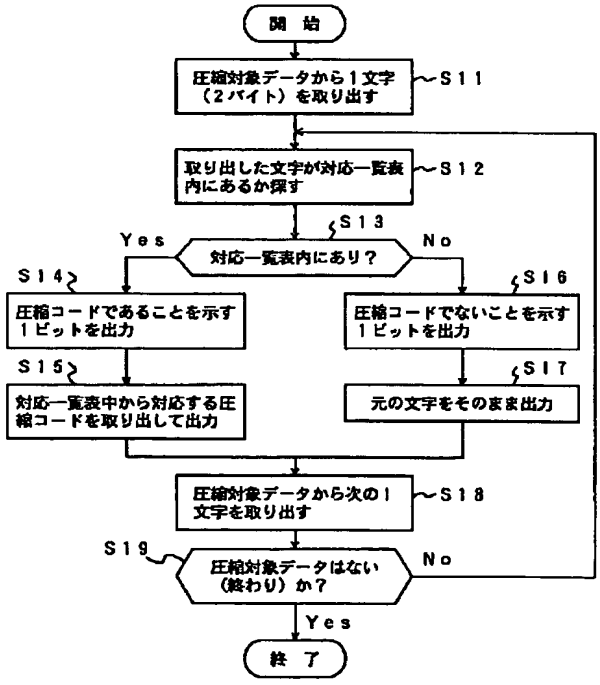
(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　戸田　亜津子
愛知県瀬戸市穴田町991番地　株式会社東芝愛知工場内

(74)代理人　弁理士　鈴江　武彦

(54)【発明の名称】　データ圧縮方法及びデータ圧縮装置

(57)【要約】
【課題】日本語のテキストデータを圧縮対象とした場合に効果的なデータ圧縮を行う。
【解決手段】特定文字のコードデータとその圧縮コードデータとを対応付けた対応一覧表を用意しておき、圧縮対象データから取り出した１文字分の文字コードデータがその対応一覧表に存在する否かを検索する（Ｓ１１，Ｓ１２）。その結果、当該データが対応一覧表に存在する場合には（Ｓ１３）、その圧縮コードデータを対応一覧表から読み出して出力し（Ｓ１４，Ｓ１５）、存在しない場合には（Ｓ１３）、当該データをそのまま出力する（Ｓ１６、Ｓ１７）。



【特許請求の範囲】
【請求項１】　特定文字のコードデータとその圧縮コードデータとを対応付けた対応一覧表を有し、
圧縮対象データが上記対応一覧表に存在するか否かを判断し、
当該圧縮対象データが上記対応一覧表に存在する場合には、その文字コードに対応する圧縮コードデータを上記対応一覧表より取り出して出力し、
当該圧縮対象データが上記対応一覧表に存在しない場合には、その文字コードデータをそのまま出力するようにしたことを特徴とするデータ圧縮方法。
【請求項２】　出力データが圧縮されたデータか否かを識別するための情報をその出力データに先立って出力するようにしたことを特徴とする請求項１記載のデータ圧縮方法。
【請求項３】　特定文字のコードデータとその圧縮コードデータとを対応付けた対応一覧表を記憶するための記憶手段と、
圧縮対象データが上記記憶手段に記憶された対応一覧表に存在するか否かを判断する判断手段と、
この判断手段により当該圧縮対象データが上記対応一覧表に存在すると判断された場合には、その文字コードに対応する圧縮コードデータを上記対応一覧表より取り出して出力し、当該圧縮対象データが上記対応一覧表に存在しない場合には、その文字コードデータをそのまま出力する出力手段とを具備したことを特徴とするデータ圧縮装置。
【請求項４】　上記出力手段は、出力データが圧縮されたデータか否かを識別するための情報をその出力データに先立って出力することを特徴とする請求項３記載のデータ圧縮装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、デジタルデータを扱うコンピュータなどのデータ処理装置に用いられるデータ圧縮方法及びデータ圧縮装置に関する。
【０００２】
【従来の技術】従来、デジタルデータを扱うコンピュータなどのデータ処理装置では、例えばハフマン符号化方式のように、１バイト単位のデータの出現頻度の違いを利用して、デジタルデータを圧縮する装置があった。
【０００３】この他にも、コンピュータが扱うデータ圧縮には様々な方法が存在するが、これらの基本的な方法は外国で考え出されたものであり、日本語の特徴を利用したものではなかった。
【０００４】
【発明が解決しようとする課題】ところで、日本語の文章の中には、平仮名、片仮名、数字、アルファベット、句読点といった文字が多く見られる。このような文字を対象としてデータ圧縮を考えた場合、上述したような方法では単に出現頻度の違いだけでデータ圧縮を行う方法であるため、効果的なデータ圧縮を行うことはできなかった。
【０００５】また、圧縮対象データの最小単位は１バイトであることから、２バイトを基本とする日本語のテキストデータには最適とはいえない。本発明は上記のような点に鑑みなされたもので、日本語のテキストデータを圧縮対象とした場合に効果的なデータ圧縮を行うことのできるデータ圧縮方法及びデータ圧縮装置を提供することを目的とする。
【０００６】
【課題を解決するための手段】本発明は、特定文字のコードデータとその圧縮コードデータとを対応付けた対応一覧表を用意しておき、圧縮対象データが上記対応一覧表に存在する場合にその文字コードデータに対応する圧縮コードデータを出力し、上記対応一覧表に存在しない場合にその文字コードデータをそのまま出力するようにしたものである。
【０００７】このような構成によれば、例えば平仮名、片仮名、数字、アルファベット、句読点といった日本語のテキストデータの中で出現頻度の高い文字を対象に、その文字コードデータとその圧縮コードデータを対応一覧表に登録しておけば、日本語の特徴を生かしたデータ圧縮を行うことができる。
【０００８】
【発明の実施の形態】以下、図面を参照して本発明の一実施形態を説明する。図１は本発明の一実施形態に係るデータ圧縮装置の構成を示すブロック図である。本装置は、デジタル化された日本語テキストデータに対して圧縮処理を行い、その結果として入力データ数よりも、出力データ数が少なくなるように変形することを目的としたシステムであり、ＣＰＵ１１、メモリ１２、補助記憶装置１３を主として構成される。
【０００９】ＣＰＵ１１は、本装置全体の制御を行うものであり、ここではデータ圧縮処理を実行する。メモリ１２は、例えばＲＯＭあるいはＲＡＭからなり、ここでは対応一覧表１２ａを記憶している。対応一覧表１２ａは、平仮名、片仮名、数字、アルファベット、句読点といった日本語のテキストデータの中で出現頻度の高い文字を対象に、その文字コードデータと圧縮コードデータとを対応付けている。
【００１０】補助記憶装置１３は、例えばフロッピーディスク装置（ＦＤＤ）あるいはハードディスク装置（ＨＤＤ）からなり、ここでは圧縮対象データ（圧縮前のデータ）と圧縮後のデータを格納する。
【００１１】図２は同実施形態における対応一覧表１２ａの内容を示す図である。対応一覧表１２ａには、平仮名、片仮名、数字、アルファベット、句読点といった日本語のテキストデータの中で出現頻度の高い文字のコードデータとそれに対応する圧縮コードデータが格納され

ている。この例では、文字コードデータは１６ビット、圧縮コードは８ビット（１６進）で表現されている。

【００１２】次に、同実施形態の動作を説明する。図３は同実施形態におけるデータ圧縮処理の動作を示すフローチャートである。ＣＰＵ１１は、まず、補助記憶装置１３などに保持されている圧縮対象データの中から１文字分（２バイト）の文字コードデータを取り出す（ステップＳ１１）。

【００１３】次に、ＣＰＵ１１はメモリ１２に記憶された対応一覧表１２ａをアクセスし、上記取り出した文字コードデータが対応一覧表１２ａ内にあるか否か検索する（ステップＳ１２）。

【００１４】ここで、当該文字コードデータが対応一覧表１２ａ内に存在した場合には（ステップＳ１３のＹｅｓ）、その圧縮コードデータを出力することになる。その際に、ＣＰＵ１１は圧縮コードデータであることを示す１ビットの情報「１」を出力し（ステップＳ１４）、次に対応一覧表１２ａの中から当該文字コードデータに対応する圧縮コードデータを読み出し、これを出力する（ステップＳ１５）。

【００１５】一方、当該文字コードデータが対応一覧表１２ａ内に存在しなかった場合には（ステップＳ１３のＮｏ）、そのまま圧縮せずに出力することになる。その際に、ＣＰＵ１１は圧縮コードデータでないことを示す１ビットの情報「０」を出力し（ステップＳ１６）、次に対応一覧表１２ａの中から当該文字コードデータに対応する圧縮コードデータを読み出し、これを出力する（ステップＳ１７）。

【００１６】以後、同様にして、圧縮対象データの中から順次文字コードデータを取り出し、これを対応一覧表１２ａと照らし合わせながら適宜圧縮して出力する（ステップＳ１８，Ｓ１９）。これにより、対応一覧表１２ａに存在する文字コードデータについては圧縮して出力することができる。

【００１７】ここで、具体的を挙げて上記処理を説明する。例えばステップＳ１１またはＳ１８で「ア」という片仮名の文字コードデータが圧縮対象データの中から取り出されたと仮定すると、処理はＳ１２→Ｓ１３→Ｓ１４→Ｓ１５と進む。

【００１８】圧縮コードを示す１ビットの情報を「１」とすると、ステップＳ１４で「１」が出力され、ステップＳ１５で「９２」（１６進）が出力される。このステップＳ１４、Ｓ１５で出力されたデータを２進数で表現すると、図４（ａ）に示すように「１１００１００１０」となる。

【００１９】また、例えばステップＳ１１またはＳ１８で「亜」という漢字の文字コードデータが圧縮対象データの中から取り出されたと仮定すると、処理はＳ１２→Ｓ１３→Ｓ１６→Ｓ１７と進む。

【００２０】圧縮コードでないことを示す１ビットの情報を「０」とすると、ステップＳ１６で「０」が出力され、ステップＳ１７で「３０２１」（１６進）が出力される。このステップＳ１６、Ｓ１７で出力されたデータを２進数で表現すると、図４（ｂ）に示すように「００１１０００００１００００１」となる。

【００２１】このように、日本語の文字単位（１６ビット）で、出現頻度の高い文字のコードデータを予め対応一覧表１２ａに登録しておき、圧縮対象データが対応一覧表１２ａに存在する場合に、これを８ビットに置き換えて出力する。これにより、例えば図５に示すように「コンピュータで扱うデータ」といった文字列データがあった場合、圧縮前に１９２ビットあったものが、圧縮後には１１６ビットに減らして出力することができる。

【００２２】

【発明の効果】以上のように本発明によれば、対応一覧表に存在する文字コードデータを圧縮コードデータに置き換えて出力するようにしたため、例えば平仮名、片仮名、数字、アルファベット、句読点といった日本語のテキストデータの中で出現頻度の高い文字を対象に、その文字コードデータと圧縮コードデータを対応一覧表に登録しておけば、日本語の特徴を生かしたデータ圧縮を行うことができる。

【００２３】この場合、基本的にＬＺ法などの適応型圧縮法と異なり、以前に出現したデータを利用しないので、対象データが短くても圧縮することができる。また、静的ハフマン符号化と異なり、対応一覧表を出力しないので、圧縮率を高められる。さらに、ハフマン符号化と異なり、データの出現頻度を算出する処理が不要であるので処理速度を高速化できる等の効果がある。

【図面の簡単な説明】

【図１】本発明の一実施形態に係るデータ圧縮装置の構成を示すブロック図。

【図２】同実施形態における対応一覧表の内容を示す図。

【図３】同実施形態におけるデータ圧縮処理の動作を示すフローチャート。

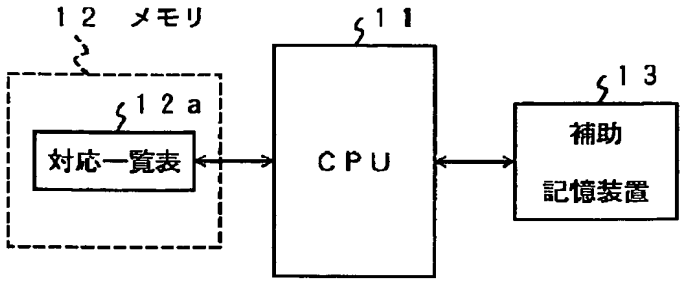
【図４】同実施形態における圧縮コードの形を説明するための図。

【図５】同実施形態におけるデータ圧縮結果を示す図。

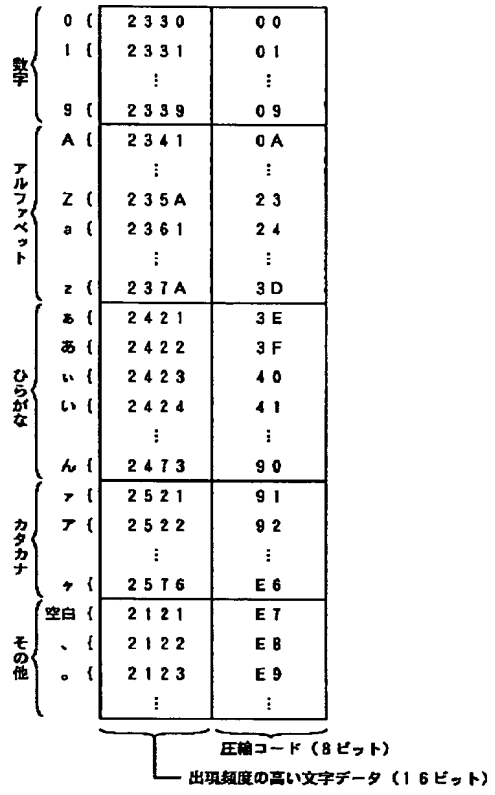
【符号の説明】

１１…ＣＰＵ、
１２…メモリ、
１２ａ…対応一覧表、
１３…補助記憶装置。

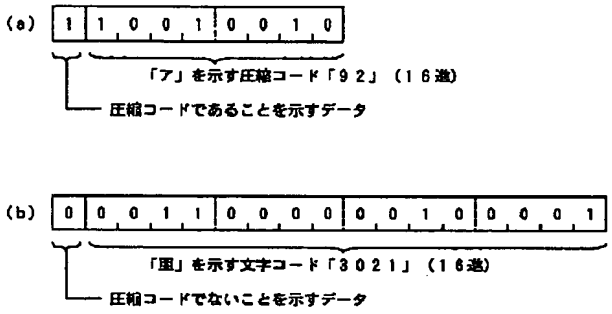
【図1】


12　メモリ
12a
対応一覧表
CPU
11
13
補助
記憶装置

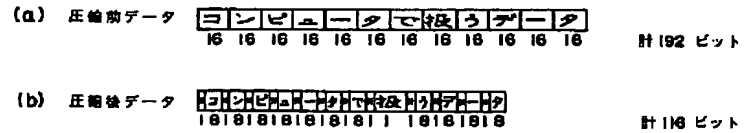
【図2】


数字
0 2330 00
1 2331 01
9 2339 09
アルファベット
A 2341 0A
Z 235A 23
a 2361 24
z 237A 3D
ひらがな
ぁ 2421 3E
あ 2422 3F
ぃ 2423 40
い 2424 41
ん 2473 90
カタカナ
ァ 2521 91
ア 2522 92
ヶ 2576 E6
その他
空白 2121 E7
、 2122 E8
。 2123 E9
圧縮コード（8ビット）
出現頻度の高い文字データ（16ビット）

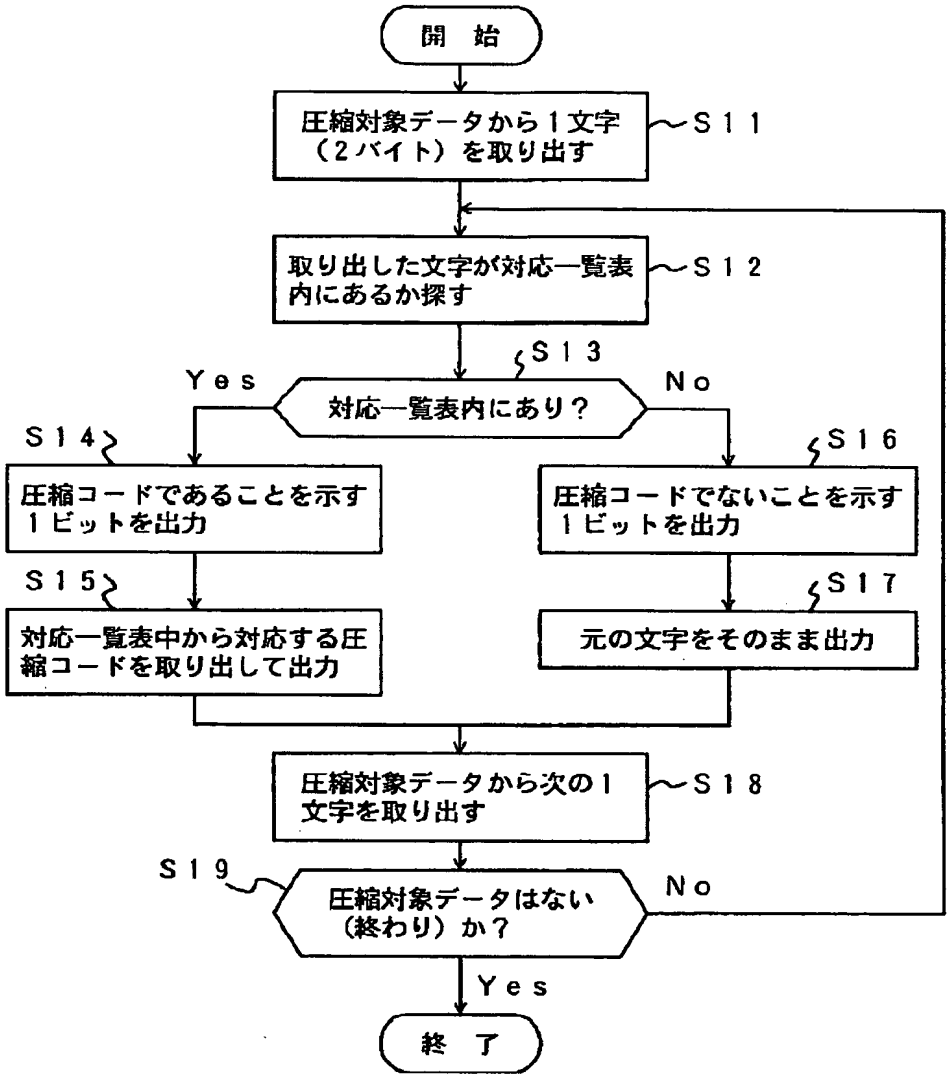
【図4】


(a)
1 1 0 0 1 0 0 1 0
「ア」を示す圧縮コード「92」（16進）
圧縮コードであることを示すデータ
(b)
0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 1
「国」を示す文字コード「3021」（16進）
圧縮コードでないことを示すデータ

【図5】


(a) 圧縮前データ
コンピュータで扱うデータ
16 16 16 16 16 16 16 16 16 16 16 16
計192 ビット
(b) 圧縮後データ
計136 ビット

【図３】


開　始
圧縮対象データから１文字（２バイト）を取り出す
S１１
取り出した文字が対応一覧表内にあるか探す
S１２
S１３
対応一覧表内にあり？
Ｙｅｓ
Ｎｏ
S１４
圧縮コードであることを示す１ビットを出力
S１５
対応一覧表中から対応する圧縮コードを取り出して出力
S１６
圧縮コードでないことを示す１ビットを出力
S１７
元の文字をそのまま出力
圧縮対象データから次の１文字を取り出す
S１８
S１９
圧縮対象データはない（終わり）か？
Ｎｏ
Ｙｅｓ
終　了